

# Representation of Molecular Structures with Persistent Homology leads to the Discovery of Molecular Groups with Enhanced CO<sub>2</sub> Binding

Jacob Townsend,<sup>†</sup> Cassie Putman Micucci,<sup>‡</sup> John H. Hymel,<sup>†</sup>

Vasileios Maroulas,<sup>‡\*</sup> Konstantinos D. Vogiatzis<sup>†\*</sup>

<sup>†</sup> *Department of Chemistry, University of Tennessee, Knoxville, TN  
37996-1600, United States*

<sup>‡</sup> *Department of Math, University of Tennessee, Knoxville, TN 37996-1320,  
United States*

E-mail: vmaroula@utk.edu; kvogiatz@utk.edu

## Abstract

Developing alternative strategies for efficient separation of CO<sub>2</sub> and N<sub>2</sub> is of general interest for the reduction of anthropogenic carbon emissions. In recent years, machine learning and high-throughput computational screening have been valuable tools in accelerated first-principles screening for the discovery of the next generation of functionalized molecules and materials. The application of machine learning for chemical applications requires the conversion of molecular structures to a machine-readable format known as a molecular representation. The choice of such representations impacts the performance and outcomes of chemical machine learning methods. Herein, we present a new concise molecular representation derived from persistent homology, an applied branch

of mathematics. We have demonstrated its applicability in a high-throughput computational screening of a large molecular database (GDB-9) with more than 133,000 organic molecules. Our target is to identify novel molecules that selectively interact with CO<sub>2</sub>. The methodology and performance of the novel molecular fingerprinting method is presented and the new chemically-driven persistence image representation is used to screen the GDB-9 database to suggest molecules and/or functional groups with enhanced properties.

The increasing concentration of greenhouse gases has been identified as a primary factor of many facets of environmental degradation such as higher global temperature, rising sea levels, increased ocean acidity, and more extreme weather-related events. CO<sub>2</sub> is the most prominent greenhouse gas, and its atmospheric concentration has exceeded 400ppm, which is more than a 40% increase from pre-industrial conditions, potentially leading to a rise in global temperatures of more than 2 °C by the year 2100.<sup>1</sup> Lowering CO<sub>2</sub> emissions is therefore mandatory to meet ambitions to limit temperature increases to 1.5 °C.<sup>2</sup> Advancements in carbon capture and storage technology are desired for meeting these goals of lowered atmospheric greenhouse gas emissions and reduced global temperature increases year to year. At an industrial level, liquid amine-based solvents are used for separation and capture of CO<sub>2</sub> via chemisorption, but the solvent regeneration step is an energy intensive process. Membrane-based technologies offer an alternative, cost-effective process for CO<sub>2</sub> capture.<sup>3-5</sup> Unlike solvents, where chemisorption involves a reaction with binding strengths exceeding 20 kcal/mol through the creation of chemical bonds between CO<sub>2</sub> and solvent, membranes utilize much weaker noncovalent interactions. Different types of materials have been suggested for the fabrication of permeable membranes including amorphous, non-porous polymeric membranes,<sup>6-12</sup> or crystalline materials with permanent porosity such as metal-organic frameworks (MOFs)<sup>13-15</sup> or zeolites.<sup>16-20</sup>

The understanding of how the atomistic structure of materials affects the gas selectivities is a crucial process for the development of more efficient carbon capture technologies. Most often, this involves the separation of CO<sub>2</sub> from N<sub>2</sub>, which are two atmospheric gases with similar kinetic diameters, making size-sieving a challenging task. The introduction of functional groups which selectively interact with CO<sub>2</sub> has been a successful approach for increasing membrane performance.<sup>12,21,22</sup> Such CO<sub>2</sub>-philic functional groups (usually Lewis bases) can be either introduced into the framework of porous crystalline materials (e.g. MOFs) or functionalized into the repeat units of non-porous polymeric membranes. Electronic structure theory calculations between molecular units and the respective gases provide a quantifica-

tion of these non-covalent interactions, as well as elucidate their nature and properties.<sup>23–29</sup> However, the number of potential CO<sub>2</sub>-philic groups is intractably large, which leads to an excessive study of such systems with accurate *ab initio* methods. Additionally, the determination of gas interaction energies may require multiple calculations to evaluate competitive gas binding sites for every structure, which further increases the computational cost and expert intervention. However, high-throughput computational screening can accelerate the discovery of new, functional materials for rational synthesis through the circumvention of the expensive and time-demanding synthesis and testing process.<sup>30–32</sup>

High-throughput computational design has shown great success in identifying new molecules<sup>33–35</sup> and materials<sup>36–41</sup> with enhanced properties and advanced functionality. For many applications, first-principles studies are essential to virtual screening, but the high computational cost of these methods makes the search of large parts of the chemical space cost-prohibitive. In recent years, machine learning (ML) has become a valuable tool in reducing the cost of a systematic chemical space exploration by enhancing the search for structure-property relationships,<sup>42–44</sup> guiding molecular design,<sup>45–48</sup> and predicting electronic structure properties.<sup>49–54</sup> ML algorithms are used for their ability to learn complicated relationships in data with high computational efficiency that can be systematically improved through additional training data, but may require extensive training set sizes before predicting out-of-sample properties accurately. The efficiency of ML depends on how these data are passed to the algorithm. For chemical applications, this occurs through molecular representations, which are the featurization of molecular compounds from their molecular structure into a vector of values. The ML algorithm then infers the relation between the structure and the property of interest. Recent developments in the formulation of molecular representations, particularly in the realm of quantum properties and structure-function relations, have increased the efficiency of ML for chemical applications.<sup>55–61</sup> In addition, such representations generalize to more complex instances such as reaction barriers.<sup>62</sup> Despite the inherent ability of ML to extract important features, ML model accuracy is dependent on the molecular

representation.

A molecular representation reduces the dimensionality of a molecular structure into a chemically meaningful format that relays important chemical information. For example, a chemical formula conveys a three-dimensional molecule as a string of characters but it is an ambiguous input for ML. Atomistic and molecular structure should be converted into a machine-readable format that can be parsed efficiently to ML algorithms as input features. One of the most prominent representations is the Coulomb Matrix (CM) introduced by Rupp *et al.*, which is a square atom-by-atom matrix containing an approximate potential energy of the free atom along the diagonal and pair Coulombic potentials on the off-diagonal terms.<sup>51</sup> An improvement over CM is typically observed using the Bag-of-Bonds (BoB) representation, where each atomic pair is placed in specific vectors (bags) based off the elemental pairs and sorted by value.<sup>52</sup> Faber *et al.* have developed FCHL, a representation based on Gaussian distribution functions for the universal kernel ridge regression-based quantum machine learning models.<sup>61</sup> Additionally, the Smooth Overlap of Atomic Positions (SOAP) representation<sup>63</sup> calculates the local density of atoms around all atoms in a given chemical environment but suffers from an increased computational cost over the pairwise CM and BoB representations.

Herein, we present a new molecular representation scheme based on *persistent homology*, a branch of computational topology. Application of persistent homology on molecules encodes three-dimensional structural data into two-dimensional *persistence images*. Since persistence images hold topological features of chemical structures, we are suggesting them as alternative molecular fingerprints. Next, those features are transposed into ML input and used to identify relationships in the data. A molecular representation is introduced for encoding chemical structures, which is applied in the prediction of interaction energies of organic molecules with gas molecules. Persistence images offer a similar-size ML vectorization regardless of system size, and a numerical example is given as a proof of this concept. Whereas CM and BoB do not provide a constant-size representation by this definition, it is

effectively achieved through padding empty cells. However, the input vector space takes the dimensions of the largest molecule in the data and requires significant padding for smaller molecules, whereas the introduced method has a predefined vector size regardless of number of atoms. It has been suggested that using feature vectors with sizes independent of system size may result in improved generalization between small and large systems.<sup>64</sup> The theoretical background of the novel molecular representation and the performance of the method in the discovery of CO<sub>2</sub>-philic functional groups is presented in the Results section. Our findings are summarized in the Discussion section, and the broader applicability of our method to chemistry is presented.

## Results

### Persistence Diagrams

The mapping of molecular structure to a chemically driven persistence image entails several steps. First, the molecular homological features, which measure the connectedness, proximity, and the empty space among the atoms, are computed and stored. These homological features are summarized in a persistence diagram (PD).<sup>65–83</sup> A PD encodes molecular features such as bonds and rings. The PDs can then be vectorized into a persistence image<sup>84</sup> (PI) for use as a molecular representation. However, employing persistent homology and its derivative, PI, purely focuses on detecting topological attributes but lacks explicit incorporation of key chemical information such as element identity, leading to limited applicability in molecular systems. Here, we describe the application of persistent homology with domain-specific knowledge for the generation of persistence images based on atomic properties. The basic steps are presented in the following paragraphs. Anisole was selected as a representative example because it contains two distinct functional units (phenyl- and methoxy- groups).

To construct a persistence diagram for a given molecule, spheres of a given radius centered at each atom are considered and, as the radius increases, the spheres intersect and lead to

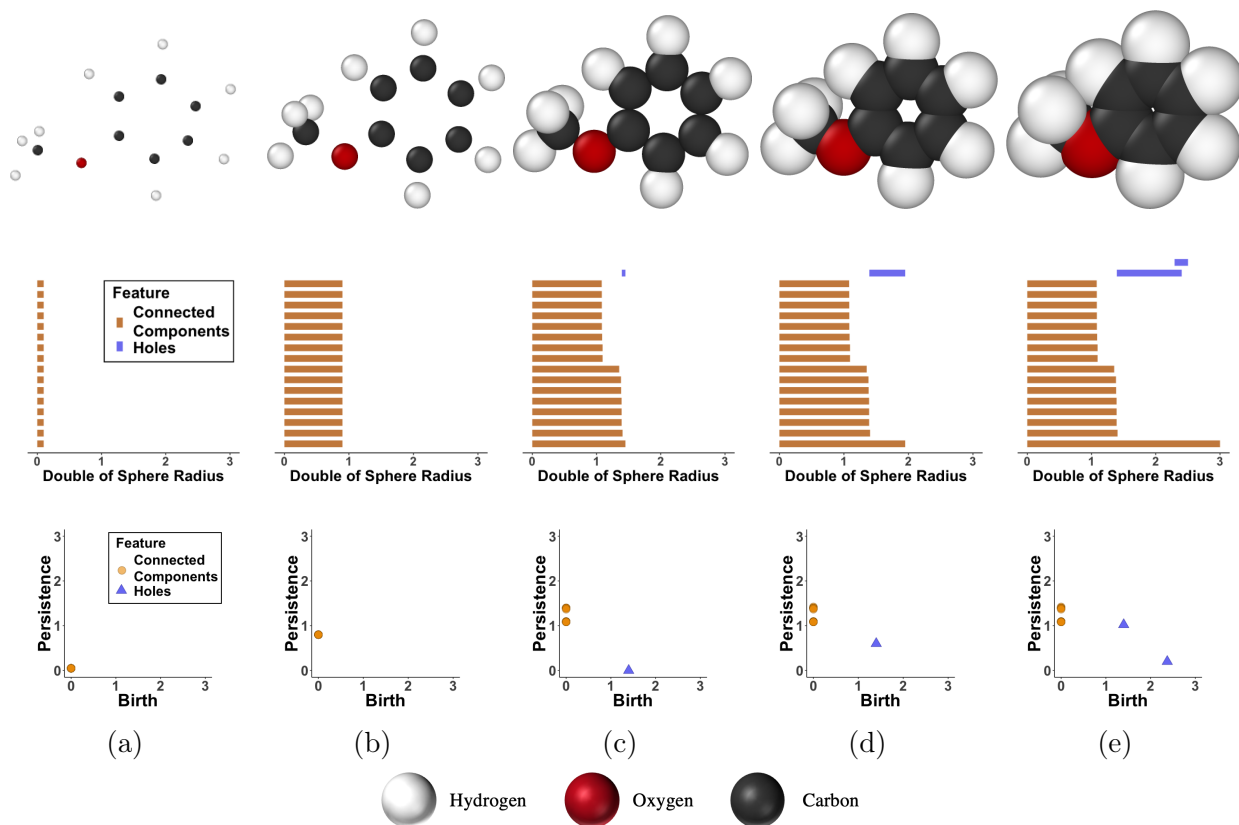


Figure 1: Graphical representation of the evolution of the PD for anisole. (a) Each atom of anisole is represented by a point (top). Each connected component is born at time  $birth = 0$  in the evolution of homological features plot (middle) and PD (bottom). Each bar at the evolution plot tracks the connected components or holes. Note that there are 16 overlapping points in the PD in (a) associated with the number of atoms. (b) The radius of a sphere centered at each atom is increased (top), and the connected components continue to persist so the persistence of a connected component is not in its final location (bottom). (c) When the radius increase leads to sphere overlap between atom pairs (top), the connected components on the PD are finalized (bottom). A hole is created by the six carbon atoms of the phenyl group (top), and a blue triangle appears in the persistence diagram to indicate its birth (bottom). (d) The radius is again expanded (top), and the hole (phenyl ring) still persists (bottom). (e) A second hole is formed between the methoxyl group and the two carbons of the phenyl group (top). The death of both holes occurs when their corresponding spheres intersect. The final PD is obtained when the death of all but one connected component and all holes are observed (bottom). Note that the bottom bar never terminates and it is always excluded from the PD.

the evolution of homological features, called *connected components* and *holes*. The connected components encode interatomic distances, while holes describe molecular attributes such as rings and functional groups.

The PDs hold information about the generation or *birth* and the lifetime length or *persistence* of connected components and holes. The placement of a birth is denoted by its location along the x-axis of the PD, whereas the *persistence* is denoted by its location on the y-axis. Birth of connected components occurs at 0, since every atom is given an initial sphere with radius 0 at the start of the algorithm (see anisole as an example on Figure 1a). The spheres are then systematically expanded (Figures 1b-e) until spherical intersections occur, which effectively generate a new connected component by merging older ones. The persistence of connected components is then recorded on the associated PD. In some sense, a PD records the time in terms of spheres’ radii for the atoms to form a single cluster. For anisole that is used here as an example, four different types of connected components are generated. The first two appear at  $\sim 1.1$  and correspond to the C-H bonds of the methoxyl and phenyl groups, respectively. The other two appear at  $\sim 1.4$  and correspond to the C-O and C-C bonds, respectively. This means that the units of the two axes are given in angstroms ( $\text{\AA}$ ). When sphere intersections lead to the formation of connected atoms (connected component) on a ring, for example when all the six spheres of the phenyl carbons have met, a hole is generated (Figure 1c). The death of a hole occurs when all spheres that form a given hole intersect, and its persistence is recorded on the persistence diagram (Figure 1e).

It becomes now evident that the connected components depend on the distances of neighboring atoms and the holes correspond to topological features of the functional groups. Two “holes” are now formed for the anisole example of Figure 1, which correspond to the phenyl and methoxy groups. These holes are unique for each respective unit and differentiate between different conformations of the same molecule and subtle differences in geometry.<sup>85–90</sup>



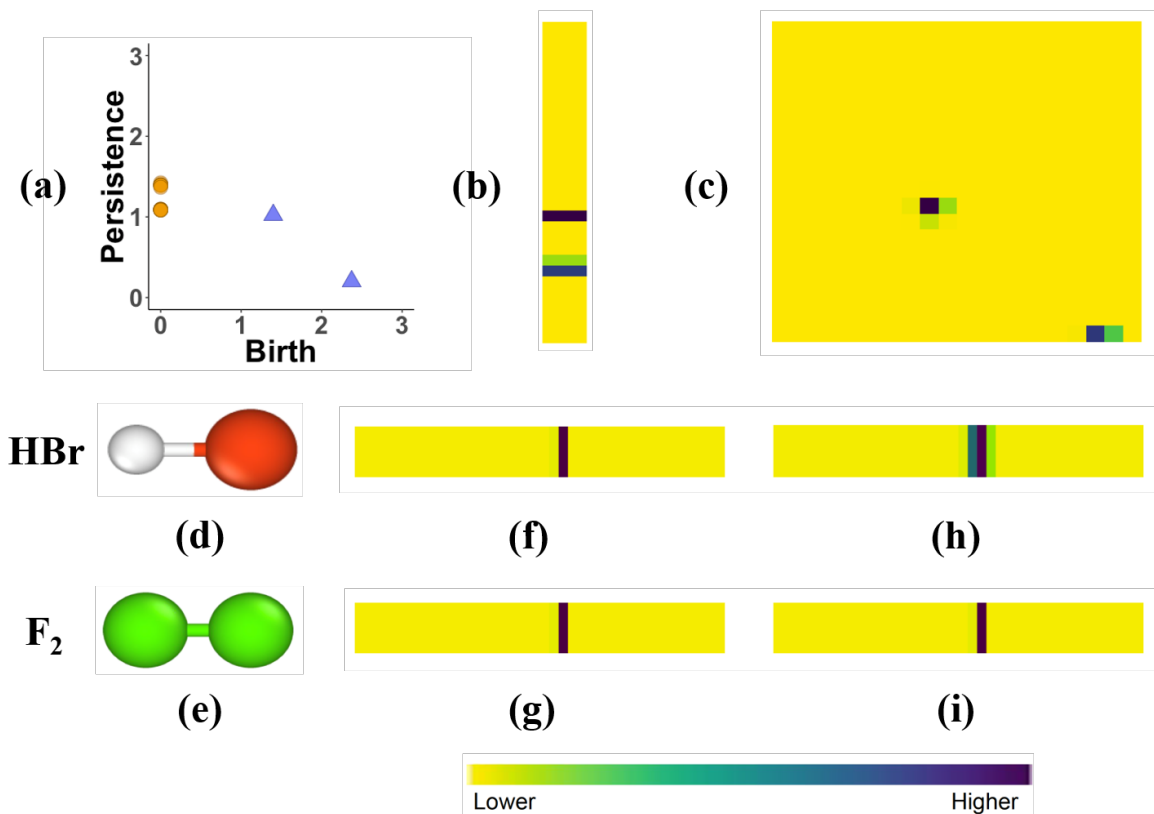


Figure 2: (a) Persistence diagram of anisole, as shown in Figure 1. The persistence image for (b) connected components and (c) holes. The molecules (d) HBr and (e)  $F_2$  have the same bond length, resulting in equivalent features on the persistence diagram. Without incorporating the chemically-driven PI, the representations in (f) and (g) are equivalent. By incorporating the pairwise electronegativity difference into the variance of the Gaussian kernel, the input vectors (h) and (i) for the two molecules are distinguished by the variance of the vector for HBr.

## Chemically-Driven Persistence Images as Molecular Representations

The PD is vectorized into a *pixelized* image, called persistence image (PI), which is a stable, computationally tractable representation.<sup>84</sup> PIs are constructed by placing a Gaussian kernel centered at each point on the PD as it is highlighted for anisole in Figure 2, where the pixel intensity corresponds to the multiplicity (Figure 2b and d). For example, five C-H pairs in the phenyl group, three C-H pairs in the methyl group etc. Next, the surface is transferred into pixel values (Figure 2c and e). The resulting image effectively encodes the molecular geometry.

The transformation of a PD to a PI may lead to inconsistencies if all atoms are treated identically, especially when molecular structures with the same geometries but different atom types are encoded into a PI. For example, the diatomic molecules HBr and F<sub>2</sub> have approximately the same bond distance (1.41 Å), and therefore generate the same PI (Figures 2g and 2k, respectively). To ameliorate this shortcoming, we introduce atomistic information in the variance of the Gaussian kernels that yielded a PI. The variance determines the spread of the kernel, or how "smeared" each point on the PD is when placed onto the PI. This variance is chosen based on the atom type that created the point in the PD. Specifically, we define the variance in the persistence images by the difference in electronegativity for connected components. Electronegativity differences are chosen because they provide a general description between the nature of different bonds. For the example of HBr and F<sub>2</sub>, HBr has a very polar chemical bond, whereas molecular fluorine is nonpolar. Large variance is provided to atom pairs with large electronegativity differences, which ultimately generates unique PIs (Figures 2h and 2l, respectively). Our new chemically-driven persistence image differentiates between molecules which have similar geometric configurations but different atomic compositions.

Another important feature is related to the dimensionality of the PI molecular representation which remains of the same order with respect to the molecular size as we show empirically in Figures S2 and S3. This is what we call herein a *similar-size* representation. For example, for a small molecule like anisole that is composed of 15 atoms, a 3Å×3Å PD was generated. The equivalent PD of a medium-size molecule such as the *tert*-butylcalixarene (105 atoms) is of comparable size (4Å×4Å). , the PD of a large structure, the main protease of the new coronavirus identified as COVID-19<sup>91</sup> in complex with an inhibitor N3 (2500 non-hydrogen atoms) has size of 6Å×6Å. A detailed analysis is given in Section S2 of the Supplementary Information. As it was mentioned in the introduction, such a similar-size representation is desirable for many chemical applications.<sup>64</sup>

## Performance of Persistence Images as Molecular Representations

Here, we demonstrate the performance of the chemically-driven PIs on an application relevant to green chemistry. Our aim is to screen a large molecular database in order to discover molecular groups that show a stronger affinity for CO<sub>2</sub> interaction over N<sub>2</sub>. Such molecular groups can be introduced in polymeric materials for the development of the next generation of functional gas separation membranes. Since it is desirable to avoid any DFT-optimized geometries as input for ML models, which introduce a significant computational bottleneck for the screening of large molecular databases, we resort to structures generated by the OpenBabel<sup>92</sup> software package (gen3d function). Our target is to train a ML model that maps low-cost geometries with accurate quantum chemical data, so it can provide reliable interaction energies for molecular species with geometries generated on-the-fly.

We tested the performance of PI as alternative molecular representations that effectively encode chemical structures. The initial subset of 100 organic molecules was used, generated based on the procedure described in Section S3 of Supplementary Information. The interaction energies of each of these structures with CO<sub>2</sub> and N<sub>2</sub> were computed by means of density functional theory (DFT). We also wanted to compare PI with the widely utilized Coulomb matrices (CM), Bag-of-Bonds (BoB), FCHL and Smooth Overlap of Atomic Positions (SOAP) representations, since each of these are produced with little computational burden and widely implemented in a number of programming libraries.<sup>93–95</sup> PI, CMs, BoBs, FCHL and SOAP representations were generated for each structure and the performance of each representation scheme was evaluated for the prediction of gas interaction energies. For each scheme, a variety of machine learning algorithms were tested, including random forest, Gaussian process regression, and kernel ridge regression. A detailed analysis of the optimization process is given in Supplementary Information, Section S6. Overall, two machine learning models were trained per molecular representation scheme, one for CO<sub>2</sub> and one for N<sub>2</sub> interaction energies. The 10-fold cross validation root mean squared error (RMSE) for the five trained models on the CO<sub>2</sub> energies are shown in Figure 3. Similar results were ob-

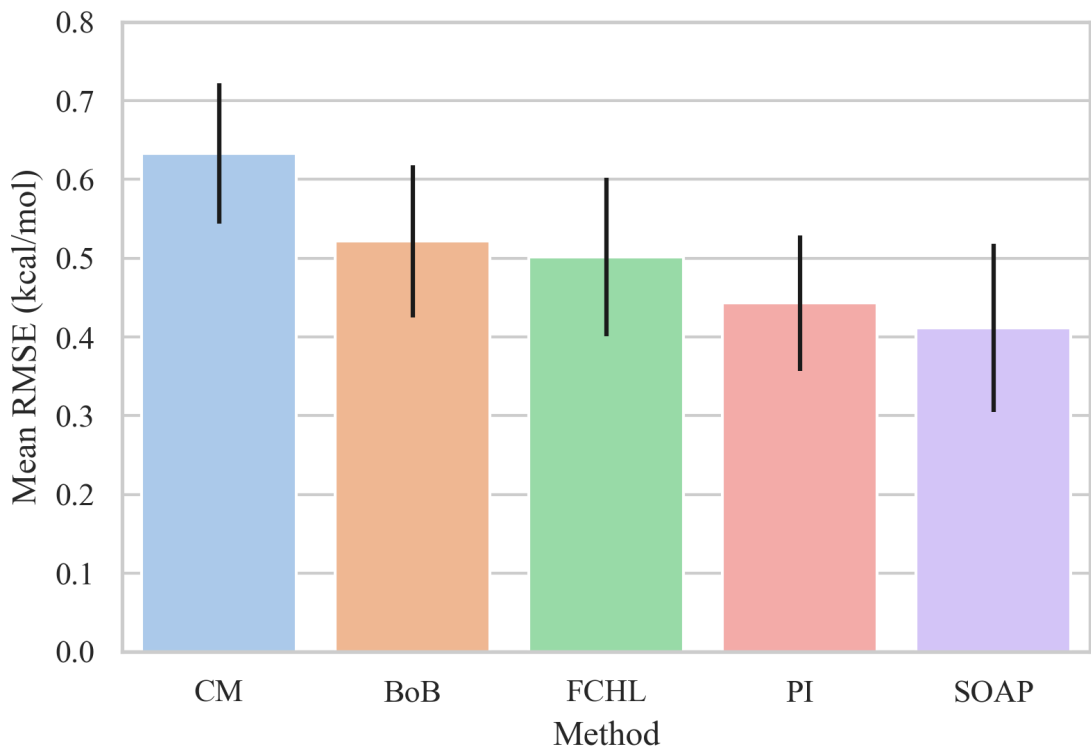


Figure 3: 10-fold cross validation root-mean-square error (RMSE, in kcal/mol) of CO<sub>2</sub> interaction energies for the CM, BoB, FCHL, PI, and SOAP molecular representation models using the automated OpenBabel-generated structures. The error bars represent the standard deviation of the RMSE.

tained for N<sub>2</sub> interaction energies (see Supplementary Information, Section S7). Comparing the best learners for each representation, CM showed the highest deviation (RMSE of 0.63 kcal/mol), followed by BoB and FCHL (RMSE of 0.52 and 0.50 kcal/mol, respectively). The most accurate models were PI with kernel ridge regression (Laplacian kernel, 0.44 kcal/mol) and SOAP with kernel ridge regression (linear kernel, 0.41 kcal/mol), where PI showed a tighter variance, yielding a higher confidence in the predictions.

## Screening the GDB-9 Database

High-throughput computational screening using ML is an efficient method to survey molecules for numerous chemical applications. Here, we are applying the PI method for identifying

molecules and functional groups that enhance CO<sub>2</sub> interactions with little computational cost. ML models trained on DFT-quality data can estimate DFT-quality results for hundreds of thousands of systems within seconds, while the explicit computation at the DFT level is a cost-prohibitive process. The GDB-9 database<sup>96</sup> was screened, which contains 133,885 organic molecules containing no more than 9 non-hydrogen atoms to determine the most promising molecules for CO<sub>2</sub> binding.

The data from the initial 100 organic molecules discussed in the previous Section do not adequately capture the properties of the chemical space spanned by the GDB-9 database. The initial training set can be considered as “biased” since it contains largely N-containing heterocycles with small functionalizations on aromatic carbons (see Supplementary Information, Section S9). For surpassing this limitation and reliably screening the full GDB-9 space, we applied a methodology known as *active learning*. In active learning, the training set is systematically expanded to capture the necessary missing physics to accurately predict for the targeted space. The top 40 molecules were selected with respect to predicted CO<sub>2</sub> interaction strength and further investigated by the MD/DFT scheme described in the computational details (Supplementary Information, Section S4). Therefore, the training set was expanded to better infer the relationship between the molecular representation and the chemical space spanning the GDB-9 database. We have repeated this processes four times by considering different yet optimized molecular representation methods (CM, BoB, SOAP, and PI). The individual steps that were followed are shown schematically in Figure 4 and analyzed in the next paragraphs.

Three iterations were performed with each representation scheme together with the optimized machine learning algorithm, as it is discussed in the previous section and in Supplementary Information, Section S6. Thus, the kernel ridge regression (Laplacian) was used for CM and PI, Gaussian process regression for BoB, and kernel ridge regression (linear) for SOAP. The active learning process resulted a total of 220 data points, i.e. 220 molecular structures with their corresponding CO<sub>2</sub> and N<sub>2</sub> interaction energies computed by DFT *per*

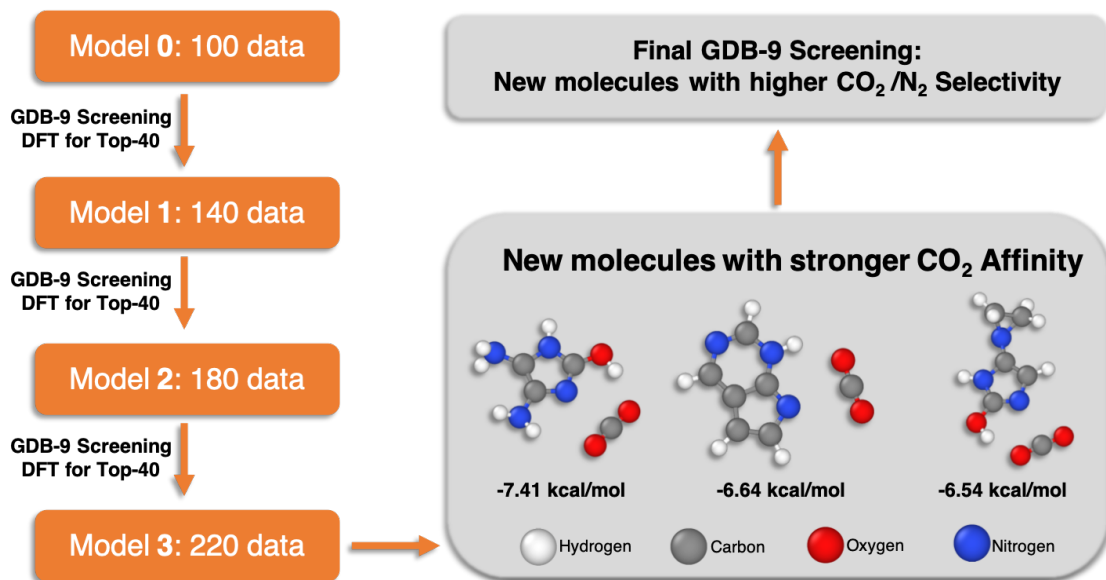


Figure 4: Flowchart of the active learning steps (left) and the results obtained from DFT (bottom, right) and the screening of the GDB-9 database (top, right).

*method.* The distributions of the interaction energies of these molecules for each method are visualized in Figure 5. We set a mark at  $-6.0$  kcal/mol for molecules with significantly strong  $\text{CO}_2$  interaction energy. For a detailed analysis of the mean and median of each active learning iteration per method, we refer the reader to the Supplementary Information, Section S8. The first iteration contains only the original 100-molecule training set. By expanding the training set with the 40 molecules from the first iteration, the next 40 best-predicted molecules from the model that utilizes the PI representation have significantly improved. On the contrary, no significant changes were observed from the other three models. By the third iteration, the  $\text{CO}_2$  distribution from CM, BoB and PI remained almost unchanged, while a small shift towards stronger interaction energies was found for the SOAP model. Overall, PI showed the greatest performance since each respective iteration increased the number of promising structures, from 10 (first iteration) to 43 (second iteration), and ultimately to 75 out of 120 molecules. Active learning with the PI molecular representation has systematically expanded the training set to better represent the chemical space of the dataset, yielding more reliable predictions every round. Since promising structures are rare

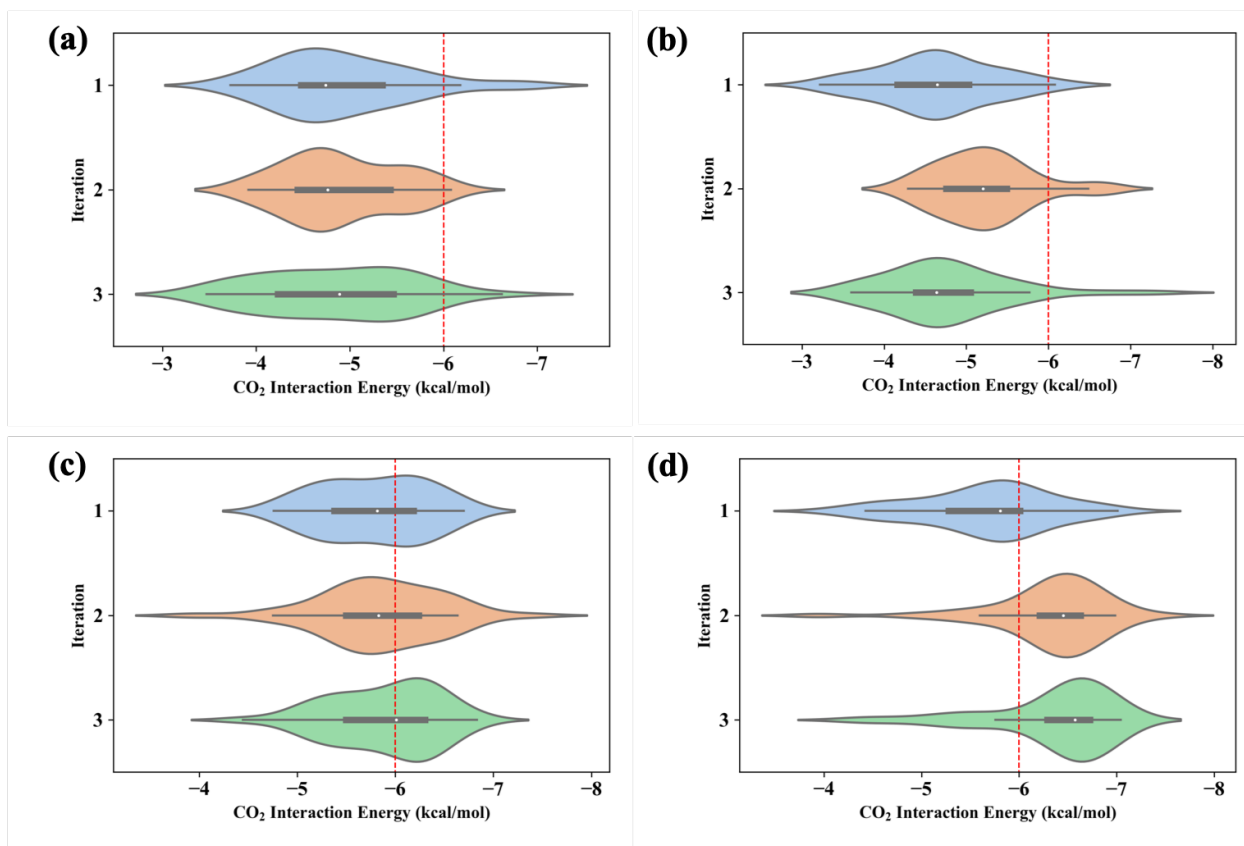


Figure 5: Horizontal violin plots showing the distributions of the CO<sub>2</sub> interaction energies for the first, second, and third iterations of the active learning process for (a) CM, (b) BoB, (c) SOAP, and (d) PI molecular representations. The height of the shape shows the frequency of occurrences at that CO<sub>2</sub> interaction energy. The white dot along the axis iteration depicts the median, and the bold line depicts the middle 50 percent of the data. The thinner line corresponds to the range of all the data. For reference, the dashed line at -6.00 kcal/mol has been placed to mark highly performing molecules.

within the dataset, this strategy allows the model to account for these rare instances within the training set in a way that would be impossible with a randomly chosen training set. Additionally, the three top candidates for CO<sub>2</sub> separations were found to demonstrate stronger interaction energies than -6.50 kcal/mol, which are shown in Figure 4 (bottom, right). Our computational procedure allowed us to discover new molecules with higher CO<sub>2</sub> affinity that combine previously unknown binding motifs. In particular, we found that cooperative effects between N-containing heterocycles with amino or hydroxo groups at ortho position increases the CO<sub>2</sub> strength. The lone electron pair of nitrogen induces a dipole moment on CO<sub>2</sub> that allows stronger interactions with hydrogen atoms of the NH<sub>2</sub>– and/or OH– functional groups.

After completing three iterations of active learning, the full dataset (220 molecules) is used to create a ML-model for predictions on the whole GDB-9 database. For comparison, the database was screened with the four different models, where each of them uses a different molecular representation method (CM, BoB, SOAP, and PI) and data generated from the corresponding active learning steps. The optimum learner for each method was used, as it was discussed in the previous Section, except for BoB, where the kernel ridge regression (linear) was applied (for a detailed discussion, see Supplementary Information, Section S8). Figure 6 includes a plot for each model, where all predicted N<sub>2</sub> and CO<sub>2</sub> interaction energies are on the x- and y-axis, respectively. Only the method that utilized the PI molecular representation was able to identify 4,5-diamino-1H-imidazol-2-ol as one of the molecules with the strongest CO<sub>2</sub> affinity as indicated with an orange dot on each plot of Figure 6. 4,5-diamino-1H-imidazol-2-ol was part of the training set introduced in the second step of the active learning process (Figure 4), and has a DFT CO<sub>2</sub> interaction energy of -7.41 kcal/mol. All methods agree that the majority of the molecular entries of the GDB-9 dataset have a mean CO<sub>2</sub> interaction energy centered between -3.0 and -4.0 kcal/mol and a N<sub>2</sub> interaction energy at -2.0 kcal/mol. Most of the molecules have predicted CO<sub>2</sub> interaction energies between -3.0 and -5.0 kcal/mol, which emphasize the difficulty in determining new molecules



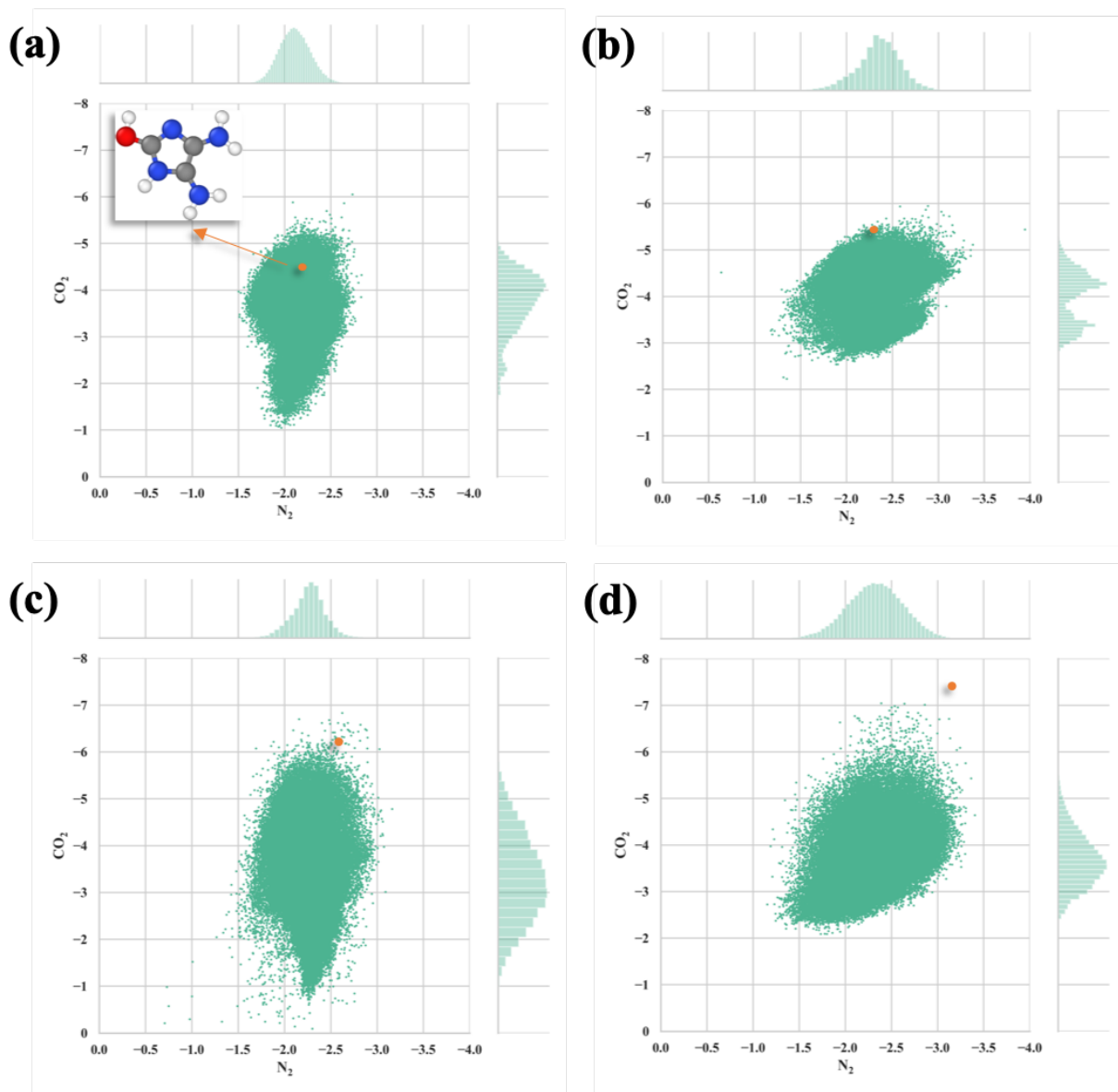


Figure 6: Predicted CO<sub>2</sub> and N<sub>2</sub> interaction energies (in kcal/mol) for all molecules in the GDB-9 database using four molecular representation models: (a) CM, (b) BoB, (c) SOAP, and (d) PI. Only the model that utilized the PI molecular representation was able to identify 4,5-diamino-1H-imidazol-2-ol (inset) as the strongest CO<sub>2</sub>-philic groups (shown with an orange dot on each plot).

with high CO<sub>2</sub> affinity. However, CM and BoB were less effective in identifying rare instances. On the contrary, results from SOAP were significantly scattered, and predicted many cases with false CO<sub>2</sub> interaction energies close to 0 kcal/mol (in a few cases, SOAP even predicted repulsive interaction energies, Supplementary Information, Section S8. Interestingly, training of machine learning models with the CM, BoB or the SOAP representations that utilize the molecules identified by active learning with the PI representation yielded more concise distributions, while all models identified the molecular species with the stronger CO<sub>2</sub> interaction energy among the top candidates (Supplementary Information, Figure S7). In other words, PI provided higher quality data for all methods. Therefore, the model that utilizes PIs for molecular representations provided the most consistent distributions for both CO<sub>2</sub> and N<sub>2</sub> interaction energies. The PI screening revealed a total of 44 of the 133,885 molecules with CO<sub>2</sub> interaction energies exceeding -6.5 kcal/mol. DFT calculations were performed for verification of these results. It should also be mentioned that SOAP needed 88,287 seconds for screening the full GDB-9 database, while the screening with the novel PI representation was almost 40 times faster (only 2,219 seconds). All screenings were performed on an Intel® i5-4278U processor.

## Discussion

A novel molecular representation utilizing persistence images with embedded chemical bonding information has been introduced for predicting DFT-quality CO<sub>2</sub>/N<sub>2</sub> interaction energies. From our investigation, this new chemically-driven persistence image is a concise, computationally efficient, and effective representation that generally outperforms other representations for prediction of CO<sub>2</sub> interaction energies since the computational cost is low and does not suffer from dimensionality problems. This representation accounts for underlying topological structure in the molecule, providing a method to control uncertainty due to differing geometric configurations. The new methodology has been applied for the screening

of the GDB-9 database to suggest new CO<sub>2</sub>-philic moieties. By using an active-learning approach, our ML-based screening was able to identify many promising molecules in the GDB-9 database despite a very small training set (220 molecules). Specifically, 44 molecules were identified that exceed -6.5 kcal/mol CO<sub>2</sub> interaction energy. Additionally, candidates that may exhibit strong CO<sub>2</sub> interactions while maintaining weak N<sub>2</sub> interactions were examined, yielding a strategy for identifying species with potentially strong gas separation capabilities. Ultimately, chemically-driven persistence images are promising molecular representations for larger supermolecular systems due to compact vectorization. Therefore, we believe that the chemically-driven PI molecular representations can be applied in a plethora of chemical problems.

The PI method described herein relies on a topological representation of a molecular compound that allows a flexible summary of the diversity of the atomic geometries. Due to this flexibility in terms of topological equivalence, the PI is robust and provides accurate predictions in contrast to other methods that need to learn rigid geometric representations. We are currently expanding the applicability of the novel molecular fingerprinting method to high-throughput screening of molecular databases for catalysis and ligand-based lanthanide/actinide separations. For this type of chemical applications, additional features are taken into consideration, such as intensity normalization when a PI is generated from the corresponding PD and predictability of properties of larger molecules from data generated from smaller ones.

## Data availability

The data for the high-throughput computational screening are available in the Supplementary Information.

## Code availability

The code for the numerical simulations is available at:

[https://gitlab.com/voglab/PersistentImages\\_Chemistry](https://gitlab.com/voglab/PersistentImages_Chemistry)

## Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant CHE-1800237 (J.T., J.H.H., and K.D.V.); and by the ARO Grant # W911NF-17-1-0313 and the NSF DMS-1821241 (CM and VM). This work used the Advanced Computer Facility (ACF) of the University of Tennessee.

## Author contributions

V. M. and K. D. V. conceived the project. J. T. and C. P. M. wrote the code. J. T. and J. H. H. performed the calculations. J. T., C. P. M., V. M. and K. D. V. wrote the paper.

## Additional information

Supplementary Information accompanies this paper at <https://doi.org/>

## Competing interests

The authors declare no competing interests.

## References

- [1] Capellán-Pérez, I., Arto, I., Polanco-Martínez, J. M., González-Eguino, M. & Neumann, M. B. Likelihood of climate change pathways under uncertainty on fossil fuel resource availability. *Energy Environ. Sci.* **9**, 2482–2496 (2016).
- [2] Hulme, M. 1.5° C and climate research after the Paris Agreement. *Nat. Clim. Chang.* **6**, 222–224 (2016).
- [3] Xu, G. *et al.* An improved CO<sub>2</sub> separation and purification system based on cryogenic separation and distillation theory. *Energies* **7**, 3484–3502 (2014).
- [4] Dutcher, B., Fan, M. & Russell, A. G. Amine-based CO<sub>2</sub> capture technology development from the beginning of 2013-A review. *ACS Appl. Mater. Interfaces* **7**, 2137–2148 (2015).
- [5] Øi, L. E. & Kvam, S. H. P. Comparison of energy consumption for different CO<sub>2</sub> absorption configurations using different simulation tools. *Energy Procedia* **63**, 1186–1195 (2014).
- [6] Norahim, N., Yaisanga, P., Faungnawakij, K., Charinpanitkul, T. & Klaysom, C. Recent Membrane Developments for CO<sub>2</sub> Separation and Capture. *Chem. Eng. Technol.* **41**, 211–223 (2018).
- [7] Mazinani, S. *et al.* A ground breaking polymer blend for CO<sub>2</sub>/N<sub>2</sub> separation. *J. CO<sub>2</sub> Util.* **27**, 536–546 (2018).
- [8] Ahmad, J. *et al.* Recent Advances in Poly (Amide-B-Ethylene) Based Membranes for Carbon Dioxide (CO<sub>2</sub>) Capture: A Review. *Polym. Technol. Mater.* **58**, 366–383 (2019).
- [9] Sekizkardes, A. K. *et al.* Microporous polymeric composite membranes with advanced film properties: pore intercalation yields excellent CO<sub>2</sub> separation performance. *J. Mater. Chem. A* 22472–22477 (2018).

- [10] Wang, Y. *et al.* Polymers of Intrinsic Microporosity for Energy-Intensive Membrane-Based Gas Separations. *Mater. Today Nano* **3**, 69–95 (2018).
- [11] Hong, T. *et al.* Carbon Dioxide Separation: Highly Permeable Oligo(ethylene oxide)-co -poly(dimethylsiloxane) Membranes for Carbon Dioxide Separation. *Adv. Sustain. Syst.* **2**, 1870030 (2018).
- [12] Hong, T. *et al.* Impact of tuning CO<sub>2</sub>-philicity in polydimethylsiloxane-based membranes for carbon dioxide separation. *J. Memb. Sci.* **530**, 213–219 (2017).
- [13] Sumida, K. *et al.* Carbon Dioxide Capture in Metal Organic Frameworks. *Chem. Rev.* **112**, 724–781 (2011).
- [14] Asgari, M. *et al.* An experimental and computational study of CO<sub>2</sub> adsorption in the sodalite-type M-BTT (M = Cr, Mn, Fe, Cu) metal-organic frameworks featuring open metal sites. *Chem. Sci.* **9**, 4579–4588 (2018).
- [15] Nguyen, H. T. *et al.* A Series of Metal-Organic Frameworks for Selective CO<sub>2</sub> Capture and Catalytic Oxidative Carboxylation of Olefins. *Inorg. Chem.* **57**, 13772–13782 (2018).
- [16] Fan, S., Liu, J., Zhang, F., Zhou, S. & Sun, F. Fabrication of zeolite MFI membranes supported by  $\alpha$ -Al<sub>2</sub>O<sub>3</sub> hollow ceramic fibers for CO<sub>2</sub> separation. *J. Mater. Res.* **28**, 1870–1876 (2013).
- [17] Kanezashi, M. & Lin, Y. S. Gas Permeation and Diffusion Characteristics of MFI-Type Zeolite Membranes at High Temperatures. *J. Phys. Chem. C* **113**, 3767–3774 (2009).
- [18] Belmabkhout, Y. & Sayari, A. Adsorption of CO<sub>2</sub> from dry gases on MCM-41 silica at ambient temperature and high pressure. 2: Adsorption of CO<sub>2</sub>/N<sub>2</sub>, CO<sub>2</sub>/CH<sub>4</sub> and CO<sub>2</sub>/H<sub>2</sub> binary mixtures. *Chem. Eng. Sci.* **64**, 3729–3735 (2009).

- [19] Krishna, R. & van Baten, J. M. Segregation effects in adsorption of CO<sub>2</sub>-containing mixtures and their consequences for separation selectivities in cage-type zeolites. *Sep. Purif. Technol.* **61**, 414–423 (2008).
- [20] Bernal, M. P., Coronas, J., Menéndez, M. & Santamaría, J. Separation of CO<sub>2</sub>/N<sub>2</sub> Mixtures Using MFI-Type Zeolite Membranes. *AIChE J.* **50**, 127–135 (2004).
- [21] Tian, Z., Saito, T. & Jiang, D.-e. Ab Initio Screening of CO<sub>2</sub>-philic Groups. *J. Phys. Chem. A* **119**, 3848–3852 (2015).
- [22] Tian, Z., Dai, S. & Jiang, D.-e. What can molecular simulation do for global warming? *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **6**, 173–197 (2016).
- [23] Vogiatzis, K. D., Mavrandonakis, A., Klopper, W. & Froudakis, G. E. Ab initio study of the interactions between CO<sub>2</sub> and n-containing organic heterocycles. *ChemPhysChem* **2**, 374–383 (2009).
- [24] Lee, H. M., Youn, I. S., Saleh, M., Lee, J. W. & Kim, K. S. Interactions of CO<sub>2</sub> with various functional molecules. *Phys. Chem. Chem. Phys.* **17**, 10925–10933 (2015).
- [25] Chen, L., Cao, F. & Sun, H. Ab initio study of the  $\pi$ - $\pi$  Interactions between CO<sub>2</sub> and benzene, pyridine, and pyrrole. *Int. J. Quantum Chem.* **113**, 2261–2266 (2013).
- [26] De Lange, K. M. & Lane, J. R. Explicit correlation and intermolecular interactions: Investigating carbon dioxide complexes with the CCSD(T)-F12 method. *J. Chem. Phys.* **134**, 034301 (2011).
- [27] Hussain, M. A., Soujanya, Y. & Sastry, G. N. Evaluating the efficacy of amino acids as CO<sub>2</sub> capturing agents: A first principles investigation. *Environ. Sci. Technol.* **45**, 8582–8588 (2011).

- [28] Townsend, J., Braunscheidel, N. M. & Vogiatzis, K. D. Understanding the Nature of Weak Interactions Between Functionalized Boranes and N<sub>2</sub>/O<sub>2</sub>, Promising Functional Groups for Gas Separations. *J. Phys. Chem. A* **123**, 3315–3325 (2019).
- [29] Hymel, J. H., Townsend, J. & Vogiatzis, K. D. CO<sub>2</sub> Capture on Functionalized Calixarenes: A Computational Study. *J. Phys. Chem. A* **123**, 10116–10122 (2019).
- [30] Kim, J., Abouelnasr, M., Lin, L. C. & Smit, B. Large-scale screening of zeolite structures for CO<sub>2</sub> membrane separations. *J. Am. Chem. Soc.* **135**, 7545–7552 (2013).
- [31] Daglar, H. & Keskin, S. Computational Screening of Metal-Organic Frameworks for Membrane-Based CO<sub>2</sub>/N<sub>2</sub>/H<sub>2</sub>O Separations: Best Materials for Flue Gas Separation. *J. Phys. Chem. C* **122**, 17347–17357 (2018).
- [32] Haldoupis, E., Nair, S. & Sholl, D. S. Finding MOFs for highly selective CO<sub>2</sub>/N<sub>2</sub> adsorption using materials screening based on efficient assignment of atomic point charges. *J. Am. Chem. Soc.* **134**, 4313–4323 (2012).
- [33] Nandy, A., Duan, C., Janet, J. P., Gugler, S. & Kulik, H. J. Strategies and Software for Machine Learning Accelerated Discovery in Transition Metal Chemistry. *Ind. Eng. Chem. Res.* **57**, 13973–13986 (2018).
- [34] Hachmann, J. *et al.* The harvard clean energy project: Large-scale computational screening and design of organic photovoltaics on the world community grid. *J. Phys. Chem. Lett.* **2**, 2241–2251 (2011).
- [35] Kanal, I. Y., Owens, S. G., Bechtel, J. S. & Hutchison, G. R. Efficient computational screening of organic polymer photovoltaics. *J. Phys. Chem. Lett.* **4**, 1613–1623 (2013).
- [36] Persson, K. A. *et al.* Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).



- [37] Levy, O. *et al.* The high-throughput highway to computational materials design. *Nat. Mater.* **12**, 191–201 (2013).
- [38] Nørskov, J. K. & Bligaard, T. The catalyst genome. *Angew. Chemie Int. Ed.* **52**, 776–777 (2013).
- [39] Collins, K. D., Gensch, T. & Glorius, F. Contemporary screening approaches to reaction discovery and development. *Nat. Chem.* **6**, 859–871 (2014).
- [40] Reetz, M. T. Combinatorial and Evolution-Based Methods in the Creation of Enantioselective Catalysts. *Angew. Chemie Int. Ed.* **40**, 284–310 (2001).
- [41] Senkan, S. Combinatorial Heterogenous Catalysis - A New Path in an Old Field. *Angew. Chemie Int. Ed.* **40**, 312–329 (2001).
- [42] Janet, J. P. & Kulik, H. J. Predicting electronic structure properties of transition metal complexes with neural networks. *Chem. Sci.* **8**, 5137–5152 (2017).
- [43] Ma, X., Li, Z., Achenie, L. E. & Xin, H. Machine-Learning-Augmented Chemisorption Model for CO<sub>2</sub> Electroreduction Catalyst Screening. *J. Phys. Chem. Lett.* **6**, 3528–3533 (2015).
- [44] Li, Z., Ma, X. & Xin, H. Feature engineering of machine-learning chemisorption models for catalyst design. *Catal. Today* **280**, 232–238 (2017).
- [45] Gómez-Bombarelli, R. *et al.* Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **4**, 268–276 (2018).
- [46] Olivecrona, M., Blaschke, T., Engkvist, O. & Chen, H. Molecular de-novo design through deep reinforcement learning. *J. Cheminform.* **9**, 1–14 (2017).
- [47] Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **361**, 360–365 (2018).

- [48] Sturluson, A., Huynh, M. T., York, A. H. P. & Simon, C. M. Eigencages: Learning a latent space of porous cage molecules. *ACS Central Science* **4**, 1663–1676 (2018).
- [49] Hansen, K. *et al.* Assessment and validation of machine learning methods for predicting molecular atomization energies. *J. Chem. Theory Comput.* **9**, 3404–3419 (2013).
- [50] Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Big data meets quantum chemistry approximations: The  $\Delta$ -machine learning approach. *J. Chem. Theory Comput.* **11**, 2087–2096 (2015).
- [51] Rupp, M., Tkatchenko, A., Müller, K.-R. & von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **108**, 058301 (2012).
- [52] Hansen, K. *et al.* Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *J. Phys. Chem. Lett.* **6**, 2326–2331 (2015).
- [53] Montavon, G. *et al.* Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.* **15**, 1–9 (2013).
- [54] De, S., Bartók, A. P., Csányi, G. & Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **18**, 13754–13769 (2016).
- [55] Faber, F., Lindmaa, A., von Lilienfeld, O. A. & Armiento, R. Crystal structure representations for machine learning models of formation energies. *Int. J. Quantum Chem.* **115**, 1094–1101 (2015).
- [56] Faber, F. A. *et al.* Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error. *J. Chem. Theory Comput.* **13**, 5255–5264 (2017).

- [57] Huang, B. & von Lilienfeld, O. A. Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity. *J. Chem. Phys.* **145** (2016).
- [58] Bereau, T., Andrienko, D. & von Lilienfeld, O. A. Transferable Atomic Multipole Machine Learning Models for Small Organic Molecules. *J. Chem. Theory Comput.* **11**, 3225–3233 (2015).
- [59] Browning, N. J., Ramakrishnan, R., von Lilienfeld, O. A. & Roethlisberger, U. Genetic Optimization of Training Sets for Improved Machine Learning Models of Molecular Properties. *J. Phys. Chem. Lett.* **8**, 1351–1359 (2017).
- [60] Bartók, A. P. *et al.* Machine learning unifies the modeling of materials and molecules. *Sci. Adv.* **3**, e1701816 (2017).
- [61] Faber, F. A., Christensen, A. S., Huang, B. & von Lilienfeld, O. A. Alchemical and structural distribution based representation for improved QML. *J. Chem. Phys.* **148**, 241717 (2018).
- [62] Meyer, B., Sawatlon, B., Heinen, S., von Lilienfeld, O. A. & Corminboeuf, C. Machine learning meets volcano plots: Computational discovery of cross-coupling catalysts. *Chem. Sci.* **9**, 7069–7077 (2018).
- [63] Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B - Condens. Matter Mater. Phys.* **87**, 1–16 (2013).
- [64] Collins, C. R., Gordon, G. J., Von Lilienfeld, O. A. & Yaron, D. J. Constant size descriptors for accurate machine learning models of molecular properties. *J. Chem. Phys.* **148**, 241718 (2018).
- [65] Bendich, P., Marron, J. S., Miller, E., Pieloch, A. & Skwerer, S. Persistent homology analysis of brain artery trees. *Ann. Appl. Stat.* **10**, 198–218 (2016).

- [66] Kramar, M., Goullet, A., Kondic, L. & Mischaikow, K. Persistence of force networks in compressed granular media. *Phys. Rev. E* **87**, 042207 (2013).
- [67] Petri, G. *et al.* Homological scaffolds of brain functional networks. *J. R. Soc. Interface* **11** (2014).
- [68] Taylor, D. *et al.* Topological data analysis of contagion maps for examining spreading processes on networks. *Nat. Comm.* **6**, 7723 (2015).
- [69] Takiyama, A., Teramoto, T., Suzuki, H., Yamashiro, K. & Tanaka, S. Persistent homology index as a robust quantitative measure of immunohistochemical scoring. *Sci. Rep.* **7**, 14002 (2017).
- [70] Li, C., Ovsjanikov, M. & Chazal, F. Persistence-based structural recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1995–2002 (2014).
- [71] Marchese, A. & Maroulas, V. Signal classification with a point process distance on the space of persistence diagrams. *Adv. Data Anal. Classif.* **12**, 657–682 (2018).
- [72] Marchese, A. & Maroulas, V. Topological learning for acoustic signal identification. In *2016 19th International Conference on Information Fusion (FUSION)*, 1377–1381 (IEEE, 2016).
- [73] Marchese, A., Maroulas, V. & Mike, J. K-means clustering on the space of persistence diagrams. In *Wavelets and Sparsity XVII*, vol. 10394, 103940W (International Society for Optics and Photonics, 2017).
- [74] Maroulas, V., Nasrin, F. & Oballe, C. A bayesian framework for persistent homology. *SIAM Journal on Mathematics of Data Science* **2**, 48–74 (2020).

- [75] Maroulas, V., Mike, J. L. & Oballe, C. Nonparametric estimation of probability density functions of random persistence diagrams. *Journal of Machine Learning Research* **20**, 1–49 (2019).
- [76] Maroulas, V., Micucci, C. P. & Spannaus, A. A stable cardinality distance for topological classification. *Advances in Data Analysis and Classification* 1–18 (2019).
- [77] Cang, Z., Mu, L. & Wei, G. W. Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PLoS Comput. Biol.* **14**, 1–44 (2018).
- [78] Cang, Z. & Wei, G. W. Analysis and prediction of protein folding energy changes upon mutation by element specific persistent homology. *Bioinformatics* **33**, 3549–3557 (2017).
- [79] Cang, Z. & Wei, G.-W. Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction. *Int. J. Numer. Meth. Bio.* **34**, e2914 (2018).
- [80] Lee, Y. *et al.* Quantifying similarity of pore-geometry in nanoporous materials. *Nat. Comm.* **8**, 15396 (2017).
- [81] Lee, Y. *et al.* High-Throughput Screening Approach for Nanoporous Materials Genome Using Topological Data Analysis: Application to Zeolites. *J. Chem. Theory Comput.* **14**, 4427–4437 (2018).
- [82] Kimura, M., Obayashi, I., Takeichi, Y., Murao, R. & Hiraoka, Y. Non-empirical identification of trigger sites in heterogeneous processes using persistent homology. *Sci. Rep.* **8**, 1–9 (2018).
- [83] Kusano, G., Fukumizu, K. & Hiraoka, Y. Kernel method for persistence diagrams via kernel embedding and weight factor. *J. Mach. Learn. Res.* **18**, 1–41 (2017).

- [84] Adams, H. *et al.* Persistence images: A stable vector representation of persistent homology. *J. Mach. Learn. Res.* **18**, 218–252 (2017).
- [85] Edelsbrunner, H. & Harer, J. Persistent homology-a survey. *Contemp. Math.* **453**, 257–282 (2008).
- [86] Zomorodian, A. & Carlsson, G. Computing persistent homology. *Discrete & Comp. Geom.* **33**, 249–274 (2005).
- [87] Ghrist, R. Barcodes: the persistent topology of data. *Bull. Am. Math. Soc.* **45**, 61–75 (2008).
- [88] Edelsbrunner, H. & Morozov, D. Persistent homology: theory and practice. Tech. Rep., Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States) (2012).
- [89] Wasserman, L. Topological data analysis. *Annu. Rev. Stat. Appl.* **5**, 501–532 (2018).
- [90] Edelsbrunner, H. & Harer, J. *Computational topology: an introduction* (American Mathematical Soc., 2010).
- [91] Jin, Z. *et al.* Structure of mpro from covid-19 virus and discovery of its inhibitors. *bioRxiv* (2020). URL <https://doi.org/10.1101/2020.02.26.964882>.
- [92] Boyle, N. M. O. *et al.* Open Babel: An open chemical toolbox. *J. Cheminform.* **3**, 1–14 (2011).
- [93] Himanen, L. *et al.* DDescribe: Library of descriptors for machine learning in materials science. *Comput. Phys. Commun.* **247**, 106949 (2020).
- [94] Christensen, A. *et al.* Qml: A python toolkit for quantum machine learning <https://github.com/qmlcode/qml>. URL <https://github.com/qmlcode/qml>.
- [95] Collins, C. R. MolML <https://github.com/crcollins/molml> (2017). URL <https://github.com/crcollins/molml>.

- [96] Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **1**, 1–7 (2014).